Predicting E-Commerce Company Success by Mining the Text of Its Publicly-Accessible Website

Dirk Thorleuchter[a,*], Dirk Van den Poel[b]

[a] Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany, dirk.thorleuchter@int.fraunhofer.de

[b] Ghent University, Faculty of Economics and Business Administration, B-9000 Gent, Tweekerkenstraat 2, Belgium, dirk.vandenpoel@ugent.be URL: http://www.crm.UGent.be

Abstract

We analyze the impact of textual information from e-commerce companies' websites on their commercial success. The textual information is extracted from web content of e-commerce companies divided into the Top 100 worldwide most successful companies and into the Top 101 to 500 worldwide most successful companies. It is shown that latent semantic concepts extracted from the analysis of textual information can be adopted as success factors for a Top 100 e-commerce company classification. This contributes to the existing literature concerning e-commerce success factors. As evaluation, a regression model based on these concepts is built that is successful in predicting the commercial success of the Top 100 companies. These findings are valuable for e-commerce websites creation.

Key Words: Success Factor, E-Commerce, SVD, Classification, Text Mining, Website

[*] *Corresponding author at: Fraunhofer INT, Appelsgarten 2, 53879 Euskirchen, Germany. Tel.: +49 2251 18305; fax: +49 2251 18 38 305*

E-mail address: Dirk.Thorleuchter@int.fraunhofer.de (D. Thorleuchter).

1 Introduction

Investigating the impact of information systems on the commercial success is a well-known topic by many researchers and practitioners (Lee & Kozar, 2006; Ballantine, Levy & Powell, 1998; DeLone & McLean, 1992; Irani & Love, 2002; Themistocleous, Irani & Love, 2004). This impact only can be measured indirectly (Galletta & Lederer, 1989) because information systems are cross-linked socio-technical entities (Serafeimidis & Smithson, 2003) with intangible benefits and indirect costs (Irani, 2002). Literature focuses this challenging task by proposing success factors that impact information system's success (DeLone & McLean, 1992; Serafeimidis & Smithson, 2003). E-commerce is a specific line of business and it differs from other line of business by the fact that the success of e-commerce companies strongly depends on companies' website quality (Lee & Kozar, 2006; Carnero, 2005; Lohse & Spiller, 1999; Ngai, 2003). Thus, specific success factors for e-commerce companies are necessary for considering these website quality aspects.

In literature, many e-commerce success factors for e-commerce companies are described (Baecke & Van den Poel, 2010). They are used to evaluate website quality (Zvirana, Glezerb & Avnia, 2006) and thus to evaluate company's success. Examples for these factors are the usability of the web page, a human computer interaction, a well-known brand, a price reduction, and a money-back guarantee. The occurrence of these factors on the company's website can be used to predict the success of the e-commerce company.

Success factors are described on companies' websites in form of semantic textual patterns and they can be identified by searching for these patterns on the website content. As example, the occurrence of a textual pattern: 'A refund will be made if you are not satisfied.' or the occurrence of a textual pattern: 'We will return your money within the first 90 days' shows that the company offers a money-back guarantee. Although both textual patterns are formulated in different ways by use of different words, they share the aspect of meaning. A semantic textual pattern includes all textual patterns with the same meaning and thus, it can be used to represent the aspect that the company offers a money-back guarantee. Further examples are that the occurrence of semantic textual patterns describing the aspect of website usage in detail or describing the aspect of interactive website services shows that the company's website is of high usability or that a human computer interaction is implemented. The occurrence of semantic textual patterns that lay great stress upon company's name or a product's name (e.g. evoked by a high frequently occurrence of specific terms) on a company's website gives a hint that company's or product's name is probably a well-known brand. As shown from these examples, the use of e-commerce success factors by a company can be identified by extracting and analyzing semantic textual patterns from company's website. However, a manual analyzing of these patterns (e.g. by human experts) from the websites of many e-commerce companies is time-consuming.

Latent semantic indexing (Christidis, Mentzas & Apostolou, 2012; Kim, Choi & Kim, 2012; Lee & Wang, 2012; Shi & Setchi, 2012; Tsai, 2012) as an automated approach can be used in this case. In contrast to further text mining approaches, it also considers semantic aspects of the texts. The approach calculates several dimensions each representing a semantic textual pattern that occurs on the websites of the e-commerce companies. Each calculated pattern can be analyzed further by comparing its aspect of meaning to the aspects of meaning of the e-commerce success factors. As a result, semantic textual patterns can be identified that represent e-commerce success factors. Thus, an automated identification of semantic textual patterns from e-commerce companies' websites representing e-commerce success factors is possible by use of latent semantic indexing.

In contrast to the existing literature concerning e-commerce success factors, we contribute two new aspects to the scientific community. The first aspect as already described above is to identify e-commerce success factors that are used by e-commerce companies based on their semantic textual patterns with latent semantic indexing. The second aspect is to evaluate the successfulness of these factors by use of a semantic textual pattern based logistic regression as modeling technique (Coussement & Van den Poel, 2008). For this second task, we extract the textual content published on the websites of the Top 100 successful e-commerce companies and of the Top 101 to 500 successful e-commerce companies separately. The identified semantic textual patterns are evaluated in terms of their usefulness for predicting most successful e-commerce companies (Top 100) in contrast to less successful e-commerce companies (Top 101 to 500). Whereas the semantic textual patterns represent success factors, this means that existing success factors for predicting successful e-commerce companies are analyzed to show their success or to show their non-success in predicting most successful e-commerce companies. These results give useful insights for e-commerce decision makers, they contribute to the existing e-commerce success factor literature, and they are valuable for e-commerce websites creation.

This work uses web mining (Thorleuchter, Van den Poel & Prinzie, 2010c) for crawling textual information from the Top 500 e-commerce companies' websites where a combined web structure mining and web content mining approach is processed. The web structure mining approach is adapted to the identification of e-commerce success factors. As an example, aspects of trustfulness are considered by crawling textual information from web pages that contain information about the awarded certifications.

In sum, the provided methodology enables the prediction of e-commerce companies' success based on information extracted from the website content of e-commerce companies. The findings in this paper give valuable insights to decision makers of e-commerce companies by identifying success factors and by evaluating the impact of the success factors on company's success. These findings also are valuable for e-commerce websites creation and they contribute to the existing e-commerce success factor literature.

2 Background

2.1 Success factors for information systems

Companies have done much investment in the procurement, implementation, and processing of information systems. These systems should increase the productivity, improve the competitiveness, and reduce operational as well as administrative costs (Molla & Licker, 2001; Schuette, 2000). Thus, decision makers of companies are interested in evaluating the success of the information systems e.g. to calculate the return on investment. Based on finding of (Galletta & Lederer, 1989) that the success of information systems can only be measured indirectly, literature shows two different ways for evaluating the success of information systems. On one hand information systems and their impact on company's success are modeled and based on the modeling results, new approaches are proposed for the evaluation (Irani, 2002; Irani & Love, 2002; Mcaulay, Doherty & Keval, 2002; Smithson & Hirschheim, 1998). On the other hand, success factors are identified that impact information system's success (DeLone & McLean, 1992; Serafeimidis & Smithson, 2003).

A well-known information system success model that is based on success factors is from DeLone and McLean (1992). The model also contains the interdependencies of the success factors and it leads to three conclusions: First, the quality of the information as well as the quality of the system itself impacts its success. Second, information systems that are easy to use and that consists of a high user satisfaction also are successful. Last, information systems with high impact on individuals and on organizational structures are more successful than other.

2.2 Success factors for e-commerce

Measuring the success of information systems by applying information system success factors is interesting for decision makers of companies. E-commerce is a specific line of business and e-commerce decision makers have done much investment in information systems in particular in their e-commerce website and they are also interested in evaluation their information systems. Literature has shown that the success of an e-commerce company strongly depends on the quality of its website (Lee & Kozar, 2006; Carnero, 2005; Lohse &

Spiller, 1999; Ngai, 2003). Thus to improve commercial success, e-commerce decision makers have the possibility to improve their website quality by considering e-commerce success factors. Specifically for e-commerce, much work has been done by researchers to identify these factors (Lee & Kozar, 2006; Baecke & Van den Poel, 2011; Delone & McLean, 1992; DeBock & Van den Poel, 2009; Lopeza & Ruiz, 2010; Lu et al., 2010; Serrano-Cinca et al., 2010; Van den Poel & Buckinx, 2005; Verhoef et al., 2010).

McKinney et al. (2002) transfer success factors from DeLone and McLean (1992) to the e-commerce domain. A high quality of the used information and a high quality of the website system itself lead to an increased internet customer satisfaction and thus, to commercial success. Devaraj et al. (2002) found that the e-commerce success is increased if customers found useful information on the website and the website itself is easy to use. Further factors are that the responding time of the website is low and that price savings are offered to the customers. Torkzadeh and Dhillon (2002) show that a wide internet product choice and that the online payment are factors that lead to an increased internet shopping convenience. Further, the internet customer relation can be improved if customers identify an internet vendor as trustful. Further factors are the shopping travel and the internet shipping that have an impact on internet shopping convenience, internet ecology, and internet product value. Unfortunately, no validation of the impact of internet shopping convenience, internet ecology, internet customer relation, and internet product value on e-commerce success is given by Torkzadeh and Dhillon (2002). Zhu and Kraemer (2002) show e-commerce success factors based on an evaluation on 260 manufacturing companies. Besides the already mentioned success factor ‘information quality’, an easy processing of the purchase also increases e-commerce success. A further factor is the customization of the website where a website is presented to each customer in an individual way. Additionally, e-commerce success can be increased by a direct website based supplier connection where goods can be delivered just in time. Besides offering high quality information on the website and besides implementing an easy to use website, Argawal and Venkatesh (2002) suggest that the information given to a customer should be made-for-the-medium. Further success factors are website promotion initiatives as well as addressing the emotions of customers. Barnes and Vidgen (2001) mention the design of a website as success factors. As further important factor the empathy with customers to ensure customer satisfaction is proposed. Koufaris (2002) point out that creating a website that increases customers shopping enjoyment is a success factor. A website also should be easy to control and the web content should be concentrate on important information. Liu and Arnett (2000) introduce the playfulness of a website in the discussion about e-commerce success factors. Loiacono et al. (2002) state that an easy to use and useful website with high entertainment effects that also consist of complementary relationships to further products is successful for e-commerce. The human computer interaction is a success measure as mentioned by Hedal (2004). Plamer (2002) proposes as e-commerce success factors the interactivity on the website, a good structured navigation, and a good organization of the website. Further, he mentioned that a small download delay is also a success factor as well as a website with quick responsiveness. Schubert (2003) shows that good website support in the four phases (information, agreement, settlement, after-sales) lead to e-commerce success. Webb and Webb (2004) focus on the reliability and the security as success factors. Wu et al. (2003) identify the cognitive outcome, the visual appearance, and the technical support as important factors for e-commerce success. Barki and Hardwick (1994) recommend focusing on the user acceptance, the user participation, the user interaction, and the user attitude to increase e-commerce success. Offering a money-back guarantee, a well-known brand, and a price reduction are mentioned by Robins et al. (2002) as an important success factor. The impact of order delivery on the success of e-commerce companies is measured by Van den Poel and Leunis (1999).

2.3 Text Classification

With text classification pre-defined classes can be assigned to textual patterns (Thorleuchter, Van den Poel & Prinzie, 2010d). Defining classes as e-commerce success factors leads to the identification of textual patterns that represent these factors. By implementing such a classification, aspects of meaning are more important than aspects of words (Thorleuchter &

Poel, 2011b). This is because textual patterns describe e-commerce success factors by use of different words. An example is the two textual patterns ‘A refund will be made if you are not satisfied’ and ‘We will return your money within the first 90 days’ as known from above. Thus, two textual patterns possibly represent the same success factor although they do not use one term in common. For a successful assignment, a text classification algorithm has to consider that the two textual pattern share no terms but they share aspects of meaning.

Literature describes many well-known text classification algorithms based on knowledge structure approaches (Palmieri & Fiore, 2010; Herranza et al., 2010). These algorithms (Support Vector Machine (SVM), Naive Bayes Classifier, Decision trees, k nearest neighbor (k-NN) classification) do not consider the aspects of meaning because underlying semantic textual patterns from a document collection are not identified.

To identify these patterns, a specific statistical technique based on a variation of eigenanalysis (eigenvectors) can be used. A semantic textual pattern does not only contain terms from one textual pattern but also further terms that are related to these terms semantically. Thus, this classification approach is able to assign e.g. success factors represented by semantic textual pattern to new textual patterns from a website even if the new patterns do not share terms with further patterns assigned to the same success factor. A well-known representative for this statistical technique is latent semantic indexing that calculates a large number of underlying semantic textual patterns automatically and that reduces their number to facilitate further processing (Thorleuchter, Van den Poel & Prinzie, 2012).

Latent semantic indexing calculates the impact of each term (e.g. words occurring on e-commerce companies' websites) on each semantic textual pattern and it also calculates the impact of each document (e.g. content of an e-commerce company's website) on the semantic textual pattern (Thorleuchter, Van den Poel & Prinzie, 2011). The calculated term impacts can be used to assign pre-defined classes (e.g. success factors) to semantic textual patterns. The calculated document impact can be used to assign document classes (e.g. the dividing of websites in two parts: in the Top 100 e-commerce companies' websites and in the Top 101 to 500 e-commerce companies' websites) to semantic textual patterns.

3 Methodology

3.1 Overview

In this paper, we use textual information from existing e-commerce companies' websites. Lists of the Top 100 and Top 500 successful e-commerce companies are used and the websites behind the companies are identified. For data collection, textual information from these websites is crawled by use of methods from web mining and is stored in documents. Documents are divided in training set and test set and they are also pre-processed by use of text mining methods. A term-website matrix based on the training set is created and latent semantic indexing is used to identify the semantic textual patterns of the training documents. To evaluate these results, the test documents are projected into the same latent semantic concept-space. A logistic regression model is built on this concept-space matrix to show that this approach is successful in predicting the most successful Top 100 e-commerce companies.

The semantic textual patterns are evaluated concerning their relation to e-commerce success factors. For this, each semantic pattern is compared to the existing success factors as described in background chapter. As a result, it possibly could be seen that some semantic textual patterns represent one or several success factors. For each of these success factors, we use the occurrence probability of its corresponding semantic pattern on a Top 100 e-commerce company's website and on a Top 101 to 500 e-commerce company's website. E-commerce success factors that mainly occur on websites of the Top 100 e-commerce companies can be used as successful classifier while success factors that mainly occur on websites of the Top 101 to 500 e-commerce companies are successful in classifying the less-successful companies. Fig. 1 shows the methodology of this approach.

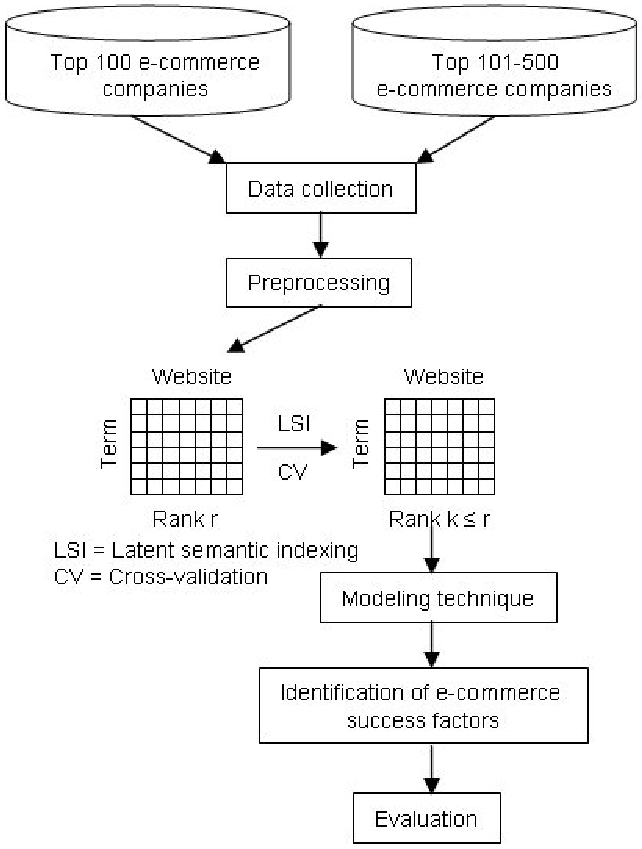
Top 100 e-commerce companies
Top 101-500 e-commerce companies
Data collection
Preprocessing
Website
Term
Rank r
LSI
CV
Website
Term
Rank k ≤ r
LSI = Latent semantic indexing
CV = Cross-validation
Modeling technique
Identification of e-commerce success factors
Evaluation

Figure 1: Different steps of the approach

3.2 Data collection

The unstructured content information from e-commerce companies' websites is collected by use of web mining methods. After identifying companies' websites, it is considered that a website consists of several web pages. Crawling all textual information from all web pages (e.g. 'sitemap', disclaimer', 'data protection policy / privacy') leads to a huge amount of information. Thus, only relevant information from companies' websites is extracted by limiting the number of web pages per company to ten. To identify the ten most relevant web pages from a company's website, the starting page is selected. Additionally, seven (sub) web pages are selected with the highest page rank returned by an internet search engine. Further relevant information might be descriptions about an e-commerce company's history. This probably indicates company's trustfulness. Additionally, relevant information also might be descriptions about its e-commerce certifications (e.g. the e-commerce certification program 'Trusted Stores' as launched by Google). Thus, web pages that contain specific terms related to a company's history or company's certifications are identified and the web pages with the highest page rank each are selected if not already selected before.

To identify the relevance of a web page concerning its page rank, Google is used as internet search engine because of the high quality of its page rank algorithm and because of the fact that all Top 500 e-commerce companies' websites can be found in the Google index. For each company, search queries are restricted to web pages of the respective company. They are automatically executed by web services (Carl, 2008) (web based advanced programming interfaces). The result data contain web pages from the company ordered by the page rank (Thorleuchter & Van den Poel, 2011a).

By use of this web mining approach, a large amount of highly unstructured information is extracted from the companies' websites. This information has to be pre-processed by use of text mining approaches to discover relevant features.

3.3 Pre-processing

To represent the extracted textual information as term vector of weighted frequencies, several methods from text mining are applied based on

- a text preparation step,
- a term filtering step,
- a vector weighting step, and
- a term vector aggregation step.

In a text preparation step, the raw text is cleaned (e.g. by deleting images, html-, or xml-tags, specific characters as well as scripting code). The punctuation is removed and a dictionary is used to correct typographical errors. Then, tokenization (Thorleuchter, Van den Poel & Prinzie, 2010a) where the term unit is word and case conversion (converting terms in lower case and capitalizing the first character) is applied.

In a term filtering step, several filtering methods (Thorleuchter, Van den Poel & Prinzie, 2010b) are used. Part-of-speech tagging is used to identify the syntactic category of a term. Stop word filtering is also used to identify terms with little or no content information (Thorleuchter, Van den Poel & Prinzie, 2008). With dictionary-based stemming, the basic form of words - where the same stem represents related words - is identified. Additionally, Zipf distribution (Zipf, 1949; Zeng, Duan, Cao & Wu, 2012) is used to reduce the number of terms by deleting rare terms. After this, the selected terms are checked manually (Gericke et al., 2009).

Then, a term vector is built. The component values of a term vector are weighted frequencies instead of using raw frequencies (the number of term-appearance in a web page) because the use of weighted frequencies significantly improves retrieval performance (Sparck Jones, 1972). Terms with large weights frequently occur in a small number of web pages but they do not occur frequently in all web pages (Salton & Buckley, 1988).

A well-known term weighting scheme is used (Salton et al., 1994) and described in formula (1). The term frequency $\mathrm{tf}_{i,j}$ equals the absolute frequency of term i in web page j, the inverse document frequency idf_i equals $\log(n/\,\mathrm{df}_i)$, the square root represents a length normalization factor, n equals the number of web pages, and m equals the dimension of the term vectors (Hotho et al., 2005).

$$\mathrm{w}_{\mathrm{i,j}} = \frac{\mathrm{tf}_{\mathrm{i,j}} \cdot \log(\mathrm{n}/\mathrm{df}_{\mathrm{i}})}{\sqrt{\sum_{\mathrm{p=1}}^{\mathrm{m}} \mathrm{tf}_{\mathrm{i,j_p}}{}^{2} \cdot (\log(\mathrm{n}/\mathrm{df}_{\mathrm{i_p}}))^{2}}} \tag{1}$$

In a term vector aggregation step, all vectors representing web pages from a specific company are aggregated to build one term vector for each company's website j (Coussement & Van den Poel, 2008). This is calculated by

$$Aw_{i,j} = \sum_{k=1}^{r} w_{i,k} \tag{2}$$

where the weight of term i in web page k is represented by $w_{i,k}$ and where r equals ten (the number of web pages per company). Then, a term-by-website matrix with weighted frequencies is created.

3.4 Semantic textual pattern identification

Normally, the term-by-website matrix is high dimensional and most of its weights are zero. To reduce the dimensionality, latent semantic indexing combined with singular value decomposition (SVD) is used. This method groups terms into concepts by forming semantic generalizations. If A is the term-by-website (m x n) matrix with rank r ($r \leq \min(m,n)$) then SVD of A is a transformation into a product of three matrices, the term-concept similarity (m x r) matrix U, the concept-website similarity (n x r) matrix V, and a diagonal (r x r) matrix Σ containing positive singular values of matrix A.

$$A = U \Sigma V^t \qquad (3)$$

To reduce the rank r of A to k ($k \leq r$), latent semantic indexing considers the first k singular values in Σ by retaining only the first k columns of U and V. Further singular values are discarded.

An important decision to be taken is the choice of k. This critical parameter influences the predictive performance. To determine the value of the parameter k, several rank k-models are constructed, are evaluated concerning their predictive cross-validated performance

(Thorleuchter, Herberz & Van den Poel, 2012), and the most favorable rank-k model is selected. To calculate the predictive performance, a prediction model is used (see Sect. 3.5).

The selected rank k-model is built on the training examples. The test examples are integrated into the same semantic subspace as created by the training examples (Deerwester, 1990).

3.5 Prediction modeling

Latent semantic indexing combined with SVD produces several semantic textual patterns standing behind the textual information provided by the content of e-commerce companies' websites. These latent semantic patterns are classified concerning their above-chance frequent occurrence in the content of the Top 100 e-commerce companies' websites (positive examples) or concerning their above-chance frequent occurrence in the content of the Top 101 to 500 e-commerce companies' websites (negative examples).

The predictive performance is measured by logistic regression as modeling technique where a maximum likelihood function is maximized (Allison, 1999; Inagaki, 2010). Advantages for the use of logistic regression are the simplicity (DeLong et al., 1988), computational speed and robustness (Greiff 1998). It can be calculated by

$$P(y=1\,|\,x)=\frac{1}{1+\exp(-(w_0+wx))} \tag{4}$$

with $T=\{(x_i,y_i)\}$ the training set, $i=\{1,2,\ldots,N\}$, $x \in R^n$ the n-dimensional input vector (a concept-website vector), w the parameter vector, w_0 the intercept, and $y_i \in \{0,1\}$ the corresponding binary target labels (company in Top 101 to 500, company in Top 100).

3.6 E-commerce success factor identification

Some of the semantic textual patterns represent e-commerce success factors while others represent semantic aspects that are not relevant for predicting e-commerce companies' success. Based on the literature review concerning existing success factors in Sect. 2, the calculated semantic textual patterns are compared to these factors. A comparison is possible because each semantic textual pattern consists of terms (words) and of the calculated impact on the semantic textual pattern. Further, each e-commerce success factor also can be described in different terms (words). As a result, human experts are able to identify e-commerce success factors standing behind a semantic textual pattern.

Based on the results of prediction modeling, we identify three groups of semantic textual patterns. Members of the first group significantly occur on the positive examples. This means, these semantic textual patterns often occur on the websites of very successful e-commerce companies (Top 100). Members of the second group significantly occur on the negative examples. These semantic textual patterns can be used to predict e-commerce companies in the Top 101 to 500 list. Further patterns (third group) can not be used for prediction. Based on the first and second group, e-commerce success factors can be identified standing behind the textual pattern of the first group. Further, e-commerce success factors for the Top 101 to Top 500 companies can be identified from the textual pattern of the second group.

3.7 Evaluation criteria

This evaluation is done with the commonly used criteria: cumulative lift, precision, recall, area under the receiver operating characteristics curve (AUC), sensitivity, and specificity. Cumulative lift measures the increase in density concerning the number of Top 100 companies relative to the density of the companies in total. This measure is most commonly used for business applications. Based on the results of prediction modeling, a list of e-commerce companies can be created sorted from most profitable to least profitable. With cumulative lift it is possible to show the density of the Top 100 companies in the top 10, top 20, or top 30 percentile of this list.

Prediction modeling uses a specific threshold to predict a company as successful or not. The precision measures the fidelity or exactness and the recall measures the completeness of the predicted results. The proportion of positive cases predicted to be positive is named sensitivity and the proportion of negative cases predicted to be negative is named specificity. To calculate these measures, the number of companies in the Top 100 list that are predicted as successful (TP) is calculated as well as the number of companies predicted as successful in the Top 101 to 500 list (FP) . Further, the number of companies in the Top 100 list that are predicted as non-successful (FN) is calculated as well as the number of companies predicted as non-successful in the Top 101 to 500 list (TN). Then, the sensitivity can be calculated by (TP/(TP+FN)), the specificity by (TN/(TN+FP)), the precision by (TP/(TP+FP)), and the recall by(TP/(TP+TN)).

It is important to know that these measures based on the selection of a specific threshold. Selecting different threshold values for the classification decision lead to different values for the performance measures. A performance measure that considers different threshold values is the AUC. It represents the area under the receiver operating characteristic curve (ROC), a two dimensional plot of two performance measures: the sensitivity versus and the inverted specificity (1-specificity). It can be used as performance measure for binary classification (Hanley & McNeil, 1982).

4 Case Study

4.1 Research data

In this study, we use lists of the Top 100 and Top 500 successful e-commerce companies as published on the internet (www.welt.de and www.internetretailer.com). The corresponding websites behind these companies are manually identified. Normally, successful companies offer websites in several languages in the internet. However here, the selected websites are restricted to the English language to prevent the language translation problem. As a result, all 500 companies offer websites in English language. Thus, 500 textual documents are created that contain content information from the most relevant websites of each company in English.

Table 1 provides summary information of the (randomly-selected) training and test set. The optimal SVD dimension is calculated using the training set and a regression model is estimated. The test set is used to show the success of the regression model compared to the frequent baseline as calculated from the relative percentage in Table 1.

	Number of customer groups	*Relative percentage*
Training set:		
Top 100 companies' websites	50	20
Top 101 to 500 companies' websites	200	80
Total	250	
Test set:		
Top 100 companies' websites	50	20
Top 101 to 500 companies' websites	200	80
Total	250	

Table 1: Overview of website characteristics

4.2 Optimal dimension selection

The result of the pre-processing step is a term-by-website matrix with high dimensionality. The training set is used to calculate an optimal SVD dimension with a cross-validation procedure (see Fig. 2). The number of concepts is represented by the x-axis and the cross-validated AUC is represented by the y-axis. The cross-validated AUC increases in the range of 2-17 concepts, it reaches a maximum at 18 concepts, and from 19 concepts on, it decreases. Thus, 18 concepts were selected as the optimal number for the SVD dimension.

The test set is integrated into the 18 semantic SVD dimension and prediction modeling is done.

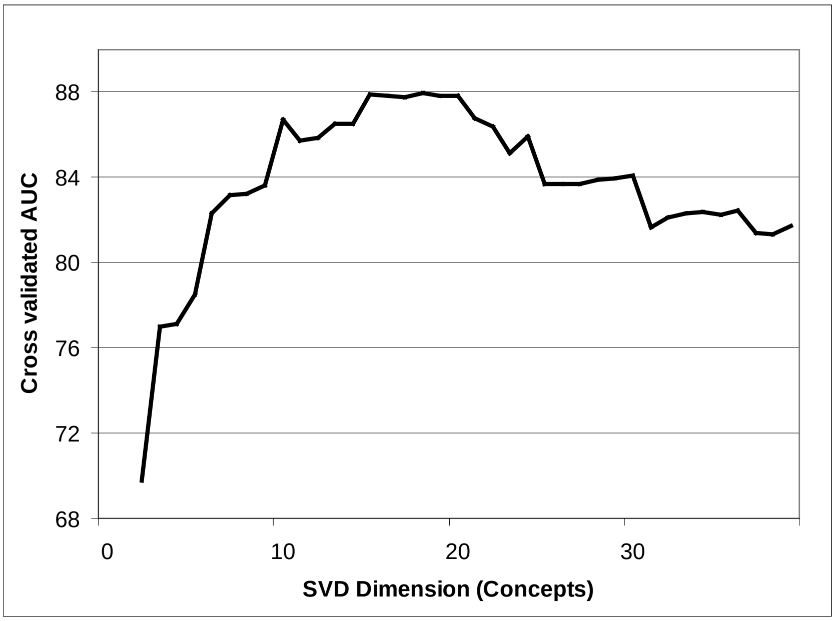

Figure 2: Calculating an optimal SVD dimension

4.3 Case study results

4.3.1 Assigning SVD dimensions to e-commerce success factors

For each of the 18 dimensions, human experts analyze the corresponding semantic patterns and compare them to terms that describe the e-commerce success factors as mentioned in

Sect. 2.2. As a result, human experts identify some semantic patterns that can be assigned to an e-commerce success measures. Further, they identify some patterns that are not related to an e-commerce success factor. Additionally, some patterns are identified that can be assigned to several e-commerce success factors. An example for this is a semantic textual pattern where the corresponding terms also can be found in success factors in the area of trusts e.g. internet vendor's reliability and trustworthiness as well as web page security. Example results in detail are presented in Sect. 4.3.3.

4.3.2 Identifying the successfulness of e-commerce success measures

After logistic regression modeling, some semantic textual patterns are identified as representative for the positive examples. An above-chance frequent occurrence of the corresponding groups of terms in text patterns from the positive examples can be seen. As a result, the identified groups of terms occur frequently in the positive examples but rarely in the negative examples. Additionally groups of terms are identified that are representative for the negative examples. These groups occur frequently in the negative examples but rarely in the positive examples.

Thus, five objectives can be shown that are representative for the positive examples. The occurrence of textual information on companies' websites describing internet customer relation by rating and providing services (Torkzadeh & Dhillon, 2002) is a good predictor for the Top 100 e-commerce companies. A second predictor is the human computer interaction (Heldal et al., 2004) based on textual information about an automatic recommendation for products, features, and services concerning user-given information.

A further predictor is the internet vendor trustworthiness (Torkzadeh & Dhillon, 2002) based on textual information describing trusted financing possibilities for products. As trustworthiness occur together with internet vendor's reliability and web page security in one semantic pattern, the two success factors (Webb & Webb, 2004) also can predict the successful e-commerce companies (Top 100).

Additionally, three objectives can be shown as predictor for the negative examples. The refund of money (money-back policy / guarantee) is mainly described on web pages of the Top 101 to 500 e-commerce company list. A second predictor is a trusted order delivery that can be monitored by the user. A third predictor is to improve internet customer relation by use of newsletters.

4.3.3 Results in detail

To show the important results in detail, groups of terms that are representative for the positive examples are presented below:

A1. Service (including services, serviced, servicing etc.) and company (including companies etc.) are two terms that occur above-chance frequently in text patterns of the positive examples together with the following terms in stemmed form: Rate, provide, product, offer, user, exclusive, ecommerce, market, etc. The terms build a semantic textual pattern that describes the rating and providing of services for products in an e-commerce market to users. This is important to quality for sale and after-sale service and it leads to an improved internet customer relation. Thus, it confirms the corresponding success factor as mentioned in Torkzadeh and Dhillon (2002).

A2. Product and finance are two terms that occur above-chance frequently in text patterns of the positive examples together with the following terms in stemmed form: Individual, offer, download, business, account, resource, asset, information, service, institute, call, competition, etc. The semantic textual pattern standing behind these terms deal with providing trusted financing possibilities for products. This is important to improve vendor legitimacy and it leads to improved internet vendor trust. Thus, it confirms the corresponding success factors mentioned in Torkzadeh and Dhillon (2002) and in Webb and Webb (2004).

A3. Search and recommendation are two terms that occur above-chance frequently in text patterns of the positive examples together with the following terms in stemmed form: System, feature, product, service, automatic, user, query, multiple, etc. The corresponding semantic pattern describes an automatic recommendation of products, features, and services based on user-given search queries. This is important to improve human computer interaction. Thus, it confirms the corresponding success factor mentioned in Plamer (2002).

Furthermore, groups of terms that are representative for the negative examples are presented below:

B1. Money and refund are two terms that occur above-chance frequently in text patterns of the negative examples (Top 101 to 500 e-commerce companies) together with the following terms in stemmed form: Price, order, replace, purchase, address, cancel, product, send, back, simplify, etc. This semantic textual pattern describes a refund of money (money-back policy of a company) as predictor for the negative examples. This means if a money-back policy of a company is mentioned on the company’s website then this company is probably not in the list of the Top 100 e-commerce companies. This contradicts research results of Robins and Kelsey (2002) where money-back guarantee is a good predictor for successful e-commerce companies. A potential reason for this could be that times have changed since 1999, money-back guarantee is now quite natural, and very successful companies do not need to mention it separately.

B2. Order and delivery are two terms that occur above-chance frequently in text patterns of the negative examples together with the following terms in stemmed form: Inbox, accurate, home, select, week, view, depart, monitor, product, etc. The corresponding semantic pattern describes a trusted order delivery that can be monitored by the user. This means if a trusted order delivery of a company is mentioned on the company's website then this company is probably not a very successful e-commerce company. This contrasts order delivery as success factor as described in Van den Poel and Leunis (1999) because for a Top 100 e-commerce company, a trusted delivery and a monitoring is also quite natural and it is also not necessary to mention it on the website.

B3: Custom and newsletter are two terms that occur above-chance frequently in text patterns of the negative examples together with the following terms in stemmed form: Mail, cost, store, product, sale, subscribe, price, registration, coupon, etc. This semantic pattern describes the internet customer relation by use of newsletters. This one-directional newsletter communication is probably dated because textual information about this internet customer relation only can be found on websites of the Top 101 to 500 e-commerce companies but not in the Top 100 e-commerce companies.

4.3.4 Comparing predictive performance

The predictive performance of the regression model is compared to the baseline by use of the following criteria: Cumulative lift curve, ROC curve, and precision/recall diagram. Fig. 3, Fig. 4, and Fig. 5 show the general success of the regression model compared to the baseline. Additionally, a three-fold cross validation is used to prevent overfitting.

Fig. 3 shows that the cumulative lift curve lies above the baseline. Thus, the density concerning the number of Top 100 companies in each percentile is greater than the density from the baseline. The ROC curve of the test sets also lies above the random baseline. Thus, the AUC of the test set (61,16) is larger than that of the baseline (50,00) with a significant improvement (χ^2=0.02 , d.f.=1, p<0.001). Additionally, the precision and recall diagram lies over the baseline at all recall values. These criteria show that the model is able to better distinguish Top 100 from Top 101 to 500 companies than the baseline.

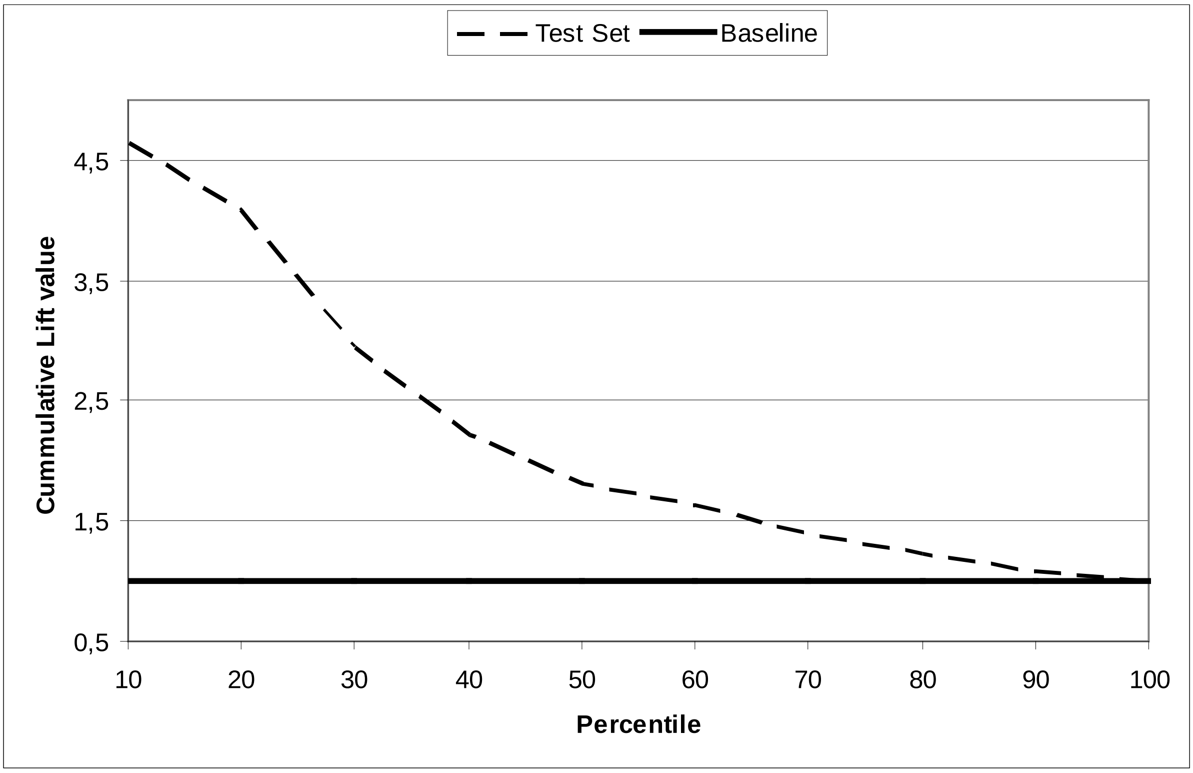

Figure 3: Cumulative lift value of the test set and of the baseline

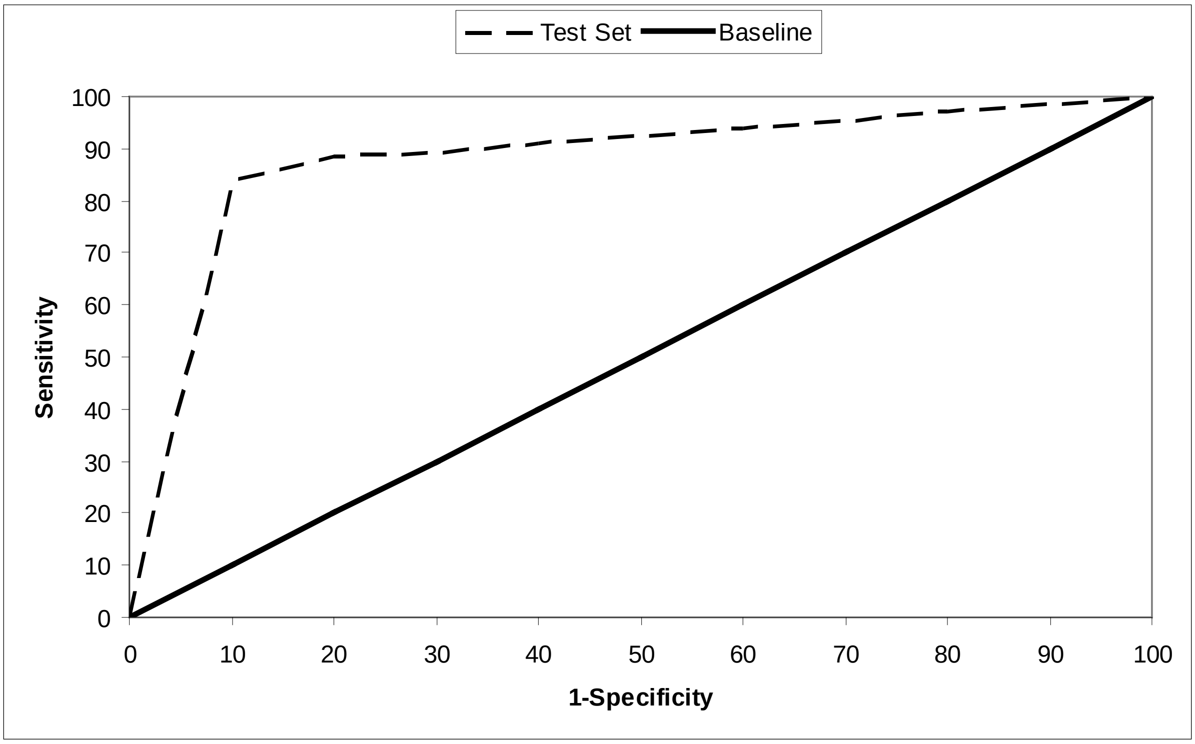

Figure 4: Sensitivity - specificity diagram of test set and baseline

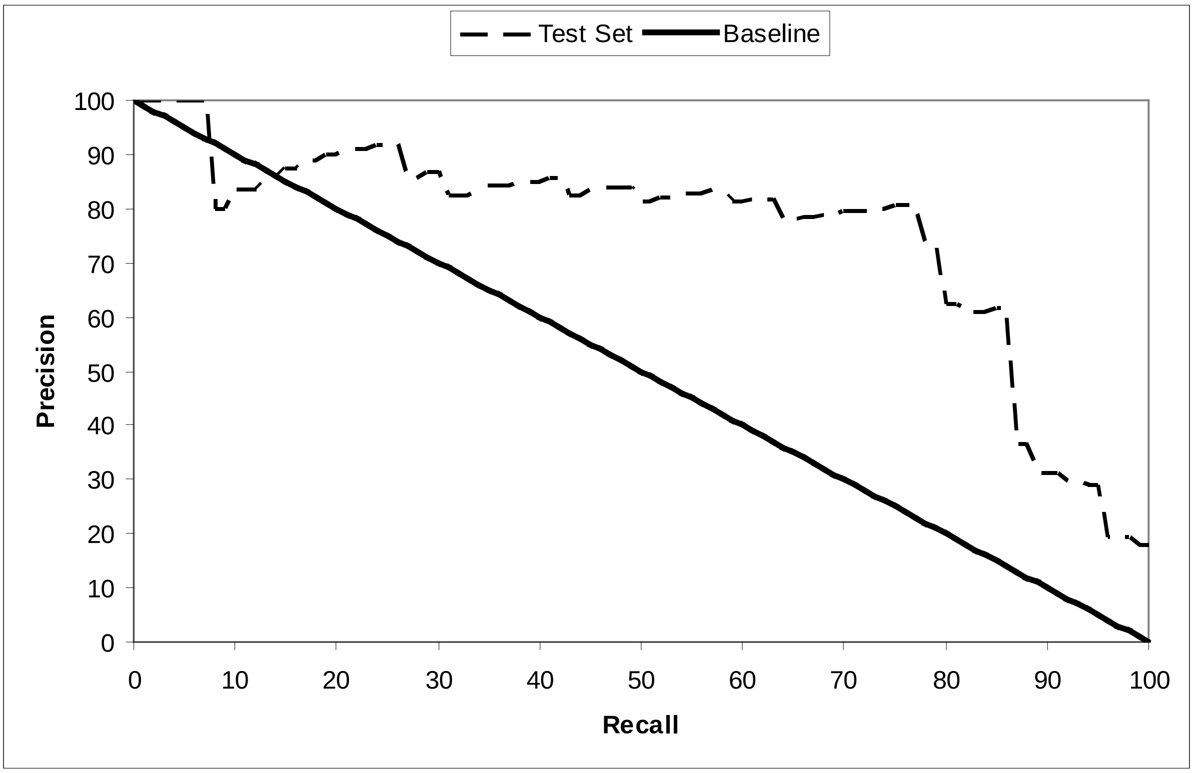

Figure 5: Precision - recall diagram of test set and baseline

5 Conclusions

This work has analyzed the impact of textual information from e-commerce companies' websites on their commercial success. Latent semantic indexing is used to identify the hidden semantic patterns on the website of the most successful Top 100 e-commerce companies and on the website of successful Top 101 to 500 e-commerce companies. Existing e-commerce success factors standing behind the semantic patterns are identified. A logistic regression model shows that predicting the most successful Top 100 e-commerce companies is successful by using the calculated semantic patterns. Thus, e-commerce success factors standing behind the semantic patterns are also successful in predicting the Top 100 companies.

Example results from the case study are that internet vendor trust, human computer interaction, and internet customer relation by rating and providing services are successful factors in predicting the Top 100 e-commerce companies and that money-back policy, trusted order delivery, and internet customer relation by use of newsletters are successful factors in predicting the Top 101 to 500 e-commerce companies.

This contributes to the existing literature concerning e-commerce success factors for e-commerce and these findings are valuable for e-commerce websites creation.

Acknowledgments

We would like to thank Joachim Schulze and Joerg Fenner for their efforts in providing us their data. We realize this project using SAS v9.1.3, SAS Text Miner v5.2, Fraunhofer Idea Web Miner v1.0, and Matlab v7.0.4. We use a self-developed program for web content mining to collect the data.

References

Allison, P. D. (1999). *Logistic Regression using the SAS System: Theory and Application.* Cary: SAS Institute Inc.

Agarwal, R. & Venkatesh, V. (2002). Assessing a firm's web presence: a heuristic evaluation procedure for the measurement of usability. Information Systems Research, 13, 168-186.

Baecke, P. H., & Van den Poel, D. (2010). Improving purchasing behavior predictions by data augmentation with situational variables. International Journal of Information Technology and Decision Making, 9(6), 853-872.

Baecke, P. H., & Van den Poel, D. (2011). Data Augmentation by Predicting Spending Pleasure Using Commercially Available External Data. Journal of Intelligent Information Systems, 367-383.

Ballantine, J., Levy, M., & Powell, P. (1998). Evaluating information systems in small and medium-sized enterprises: issues and evidence. European Journal of Information Systems, 7, 241– 251.

Barki, H., & Hardwick, J. (1994). Measuring user participation, user involvement, and user attitude. MIS Quarterly, 18(1), 59-79.

Barnes, S.J., & Vidgen, R. (2001). An evaluation of cyber-bookshops: the webQual method. International Journal of Electronic Commerce, 6, 11– 30.

Carnero, M.C. (2005). Selection of diagnostic techniques and instrumentation in a predictive maintenance program: a case study. Decision Support Systems, 38(4), 539– 555.

Carl, D., Clausen, J., Hassler, M., & Zund, A. (2008). *Mashups programmieren* (pp. 51-53). Köln: O'Reilly.

Coussement, K., & Van den Poel, D. (2008). Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Information & Management,* 45, 164-174.

Christidis, K., Mentzas, G., & Apostolou, D. (2012). Using latent topics to enhance search and recommendation in Enterprise Social Software. Expert Systems with Applications, 39(10), 9297-9307.

DeBock, K. W., & Van den Poel, D. (2009). Predicting website audience demographics for web advertising targeting using multi-website clickstream data. *Fundamenta Informaticae*, 97, 1-19.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science,* 41(6), 391-407.

Delone, W.H., & McLean, E.R. (1992). Information systems success: the quest for the dependent variable. *Information Systems Research,* 3(1), 60-95.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics,* 44(3), 837-845.

Devaraj, S., Fan, M., & Kohli, R. (2002). Antecedents of B2C channel satisfaction and preference: validating e-commerce metrics," Information Systems Research, 13, 316– 333.

Galletta, D.F., & Lederer, A.L. (1989). Some cautions on the measurement of user information satisfaction. *Decision Sciences*, 20, 419-438.

Gericke, W., Thorleuchter, D., Weck, G., Reiländer F., & Loß, D. (2009). Vertrauliche Verarbeitung staatlich eingestufter Information - die Informationstechnologie im Geheimschutz. *Informatik Spektrum,* 32(2), 102-109.

Greiff, W. R. (1998). A theory of term weighting based on exploratory data analysis. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21st SIGIR Conference* (pp. 11-19). New York: ACM.

Hanley, J.A., & McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* 143(1), 29-36.

Heldal, F., Sjøvold, E., & Heldal, A.F. (2004). Success on the Internet—optimizing relationships through the corporate site. *International Journal of Information Management,* 24(2), 115-129.

Herranza, J., Matwin, S., Nind, J., & Torra, V.,(2010). Classifying data from protected statistical datasets. Computer and Security, 29(8), 875-890.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *LDV Forum*, 20(1), 19-26.

Inagaki, S. (2010). The Effects of Proposals for Basic Pension Reform on the Income Distribution of the Elderly in Japan. *The Review of Socionetwork Strategies*, 4(1), 1-16.

Irani, Z. (2002).Information systems evaluation: navigating through the problem domain," Information and Management, 40, 11– 24.

Irani, Z. & Love, P.E.D. (2002). Developing a frame of reference for exante IT/IS investment evaluation. European Journal of Information Systems, 11, 74–82.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11-21.

Kim H.K., Choi, I.Y., & Kim, J.K. (2012). A literature review and classification of recommender systems research. Expert Systems with Applications, 39(1), 10059-10072.

Koufaris, M. (2002). Applying the technology acceptance model and flow theory to online customer behavior. Information Systems Research, 13, 205-223.

Lee, C.H., & Wang S.H. (2012). An information fusion approach to integrate image annotation and text mining methods for geographic knowledge discovery. Expert Systems with Applications, 39(10), 8954-8967.

Lee, Y., & Kozar, K.A. (2006). Investigating the effect of website quality on e-business success: An analytic hierarchy process (AHP) approach. Decision Support Systems, 42, 1383–1401.

Liu, C., & Arnett, K.P. (2000). Exploring the factors associated with web site success in the context of electronic commerce. Information and Management, 38, 23– 33.

Lohse, G.L., & Spiller, P. (1999). Internet retail store design: how the user interface influences traffic and sales. Journal of Computer Mediated Communication, 5.

Loiacono, E.T., Chen, D.Q., & Goodhue, D.L. (2002). WebQualk revisited: predicting the intent to reuse a website. In proceedings of 8th Americas Conference on Information Systems (pp. 301-309).

Lopeza, I., & Ruiz, S. (2010). Explaining website effectiveness: The hedonic–utilitarian dual mediation hypothesis. Electronic Commerce Research and Applications, 10(1), 49-58.

Lu, Y., Zhao, L., & Wang, B. (2010). From virtual community members to C2C e-commerce buyers: Trust in virtual communities and its effect on consumers' purchase intention. *Electronic Commerce Research and Applications*, 9(4), 346-360.

Mcaulay, L., Doherty, N., & Keval, N. (2002). The stakeholder dimension in information systems evaluation. Journal of Information Technology, 17, 241–255.

McKinney, V., Yoon, K., & Zahedi, F.M. (2002). The measurement of webcustomer satisfaction: an expectation and disconfirmation approach. Information Systems Research, 13, 296– 315.

Molla, A., Licker, P.S. (2001). E-commerce systems success: an attempt to extend and respecify the DeLone and McLean model of IS success. Journal of Electronic Commerce Research, 2, 131–141.

Ngai, E.W.T. (2003). Selection of web sites for online advertising using the AHP. Information and Management, 40, 233– 242.

Palmer, J.W. (2002). Web site usability, design, and performance metrics. Information Systems Research, 13, 151-167.

Palmieri F., & Fiore, U. (2010). Network anomaly detection through nonlinear analysis. Computer Security, 29, 737-55.

Robins, D., & Kelsey, S. Analysis of Web-based information architecture in a university library: navigating for known items. *Information Technology and Libraries*, 21(4), 158-169.

Salton, G., Allan, J., & Buckley, C. (1994). Automatic structuring and retrieval of large text files. *Communications of the ACM,* 37(2), 97-108.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management,* 24(5), 513–523.

Schubert, P. (2003). Extended web assessment method (EWAM): evaluation of electronic commerce applications from the customer's viewpoint. International Journal of Electronic Commerce, 7, 51–80.

Schuette, D. (2000). Turning e-business barriers into strengths. Information Systems Management, 20–25.

Serafeimidis, V., & Smithson, S. (2003). Information systems evaluation as an organizational institution—experience from a case study. Information Systems Journal, 13, 251– 274.

Serrano-Cinca, C., Fuertes-Callén, Y., & Gutiérrez-Nieto, B. (2010). Internet positioning and performance of e-tailers: An empirical analysis. *Electronic Commerce Research and Applications*, 9(3), 237-248.

Shi, L., & Setchi, R. (2012). User-oriented ontology-based clustering of stored memories. Expert Systems with Applications, 39(10), 9730-9742.

Smithson, S., & Hirschheim, R. (1998). Analysing information systems evaluation: another look at an old problem. European Journal of Information Systems, 7, 158–174.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. J Doc, 28(1), 11-21.

Themistocleous; M.; Irani; Z.; & Love, P.E.D. (2004). Evaluating the integration of supply chain information systems: a case study. European Journal of Operational Research, 159(2), 393– 405.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2012). Analyzing existing customers’ websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. Expert Systems with Applications, 39(3), 2597-2605.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2011). LSI based Profitability Prediction of new Customers," In Proc. SIAM International Workshop on Data Mining for Marketing (pp. 62-67). New York: SIAM.

Thorleuchter, D., & Van den Poel, D. (2011a). Companies Website Optimising concerning Consumer's searching for new Products. In Proc. Uncertainty Reasoning and Knowledge Engineering (pp. 40-43). New York: IEEE.

Thorleuchter, D., & Van den Poel, D. (2011b). Semantic Technology Classification - A Defence and Security Case Study. In Proc. Uncertainty Reasoning and Knowledge Engineering (pp. 36-39), New York: IEEE.

Thorleuchter, D., Herberz, S., & Van den Poel, D. (2012). Mining Social Behavior Ideas of Przewalski Horses. Lecture Notes in Electrical Engineering, 121, 649-656.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010a). Mining Ideas from Textual Information. *Expert Systems with Applications,* 37(10), 7182-7188.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010b). A compared R&D-based and patent-based cross impact analysis for identifying relationships between technologies. *Technological Forecasting and Social Change,* 77(7), 1037-1050.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010c). Extracting consumers needs for new products - A web mining approach. In *Proceedings WKDD 2010* (p. 441.). Los Alamitor: IEEE Computer Society.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010d). Mining innovative ideas to support new product research and development. In H. Locarek-Junge, & C. Weihs (Eds.),,*Classification as a Tool for Research,* Berlin: Springer.

Thorleuchter, D. (2008). Finding Technological Ideas and Inventions with Text Mining and Technique Philosophy. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning, and Applications* (pp. 413-420). Berlin: Springer.

Torkzadeh, G., & Dhillon, G. (2002). Measuring factors that influence the success of Internet commerce. Information Systems Research, 13, 87-204.

Tsai, H.H. (2012). Global data mining: An empirical study of current trends, future forecasts and technology diffusions. Expert Systems with Applications, 39(9), 8172-8181.

Van den Poel, D., & Buckinx, W. (2005). Predicting Online-Purchasing Behavior. *European Journal of Operational Research,* 166(2), 557-575.

Van den Poel, D., & Leunis, J. (1999). Consumer Acceptance of the Internet as a Channel of Distribution. *Journal of Business Research,* 45(3), 249-256.

Verhoef, P. C., Venkatesan, R., McAlister, L., Malthouse, E. C., Krafft, M., & Ganesan, S. (2010). CRM in Data-Rich Multichannel Retailing Environments: A Review and Future Research Directions. *Journal of Interactive Marketing,* 24(2), 121-137.

Webb, H.W., & Webb, L.A. (2004). SiteQual: an integrated measure of web site quality. Journal of Enterprise Information Management, 17, 430–440.

Wu, F., Mahajan, V., & Balasubramanian, S. (2003). An analysis of e-business adoption and its impact on business performance. Journal of the Academy of Marketing Science, 13, 425– 447.

Zeng, J., Duan, J., Cao, W., & Wu C. (2012). Topics modeling based on selective Zipf distribution. Expert Systems with Applications, 39(7), 6541-6546.

Zhu, K., & Kraemer, K. (2002). E-commerce metrics for net-enhanced organizations: assessing the value of e-commerce to firm performance in the manufacturing sector. Information Systems Research, 13, 275–295.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort.* Cambridge: Addison-Wesley.

Zvirana, M., Glezerb, C., & Avnia, I. (2006). User satisfaction from commercial web sites: The effect of design and use. *Information & Management,* 43(2), 157-178.